DNN-BASED SPEECH MASK ESTIMATION FOR EIGENVECTOR BEAMFORMING

Lukas Pfeifenberger[1], Matthias Zöhrer[1], Franz Pernkopf[1]

[1] Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria
lukas.pfeifenberger@alumni.tugraz.at
{matthias.zoehrer,pernkopf}@tugraz.at

ABSTRACT

In this paper, we present an optimal multi-channel Wiener filter, which consists of an eigenvector beamformer and a single-channel postfilter. We show that both components solely depend on a speech presence probability, which we learn using a deep neural network, consisting of a deep autoencoder and a softmax regression layer. To prevent the DNN from learning specific speaker and noise types, we do not use the signal energy as input feature, but rather the cosine distance between the dominant eigenvectors of consecutive frames of the power spectral density of the noisy speech signal. We compare our system against the BeamformIt toolkit, and state-of-the-art approaches such as the front-end of the best system of the CHiME3 challenge. We show that our system yields superior results, both in terms of perceptual speech quality and classification error.

Index Terms— multi-channel speech enhancement, eigenvector beamforming, speech mask estimation

1. INTRODUCTION

In recent years, conventional single channel speech enhancement methods have been outperformed by data-driven approaches. *deep neural networks* (DNNs) have been employed to discriminatively learn a gain mask for separation of the speech and noise components in a noisy speech signal [1–5].

For multi-channel speech enhancement, acoustic beamforming still outperforms single-channel methods due to the underlying physical model that can be exploited [6]. However, DNNs have proven to be useful for learning a postfilter subsequent to a beamformer [7]. The *generalized sidelobe canceller* (GSC) is one of the most popular beamformer designs. It requires an estimate of either the *direction of arrival* (DOA) or the *acoustic transfer function* (ATF) from the speech source to the microphones, which is then used as steering vector [6]. For DOA estimation, the geometry of the microphone array has to be known, while ATF estimation requires knowledge of the statistics of the speech signal. More advanced beamforming techniques require an estimate of the *power spectral density* (PSD) matrix of the noise signal [8].

In this paper, we first show that the speech presence probability mask for estimating the speech and noise statistics is sufficient to construct an optimal multi-channel Wiener filter, consisting of an *eigenvector GSC* (EV-GSC) and a single-channel postfilter. Recently, various works have been presented on how to obtain the speech presence probability using neural networks, e.g. [1, 3, 9]. Most methods rely on the energy of the noisy speech signals, and therefore are highly dependent on the array geometry and the statistics of the speech and noise presented in the training data. We aim to use a more general approach, which requires as little assumptions about the signals as possible: We only assume that the speaker is *moving slowly*, and that the noise is *non-stationary*. We empirically observed that the eigenvectors of the PSD matrix of the noisy speech signals provide a good measure for speaker activity, independent of signal energy and array geometry. Based on this observation, we estimate the speech presence probability mask using a simple DNN structure consisting of a deep autoencoder with a softmax regression layer. The deep autoencoder learns a sparse representation of the eigenvectors of the PSD matrix of the noisy speech signals for each frequency bin. The softmax regression layer discriminatively maps this representation to the speech presence probability mask. We empirically compare our multi-channel speech enhancement system to three state-of-the-art approaches: The BeamformIt-toolkit [10], a GSC with steering vector estimation and an *adaptive blocking matrix* (ABM) [7], and the front-end of the best CHiME3 system [11], which uses a complex Gaussian mixture model (CGMM-EM) to estimate the speech and noise statistics.

This paper is structured as follows: After the introduction of the system model in Section 2 we show the importance of the speech presence probability for constructing an optimal multi-channel Wiener filter in Section 3. In Section 4 the estimation of the speech presence probability is presented. In Section 5 we evaluate our model on CHiME4 data. Section 6 concludes the paper.

2. SYSTEM MODEL

We use the CHiME4 setup [10], which provides multi-channel recordings of a single speaker embedded into ambient noise. The recordings have been made with $M = 6$ microphones mounted to a tablet computer. Both real and simulated data is provided, as well as a ground truth (i.e. speaker separated from noise). This allows to evaluate the performance of our system based on the true speech signal. According to this scenario, the signal model is given as

$$\boldsymbol{Z}(k,l) = \boldsymbol{S}(k,l) + \boldsymbol{N}(k,l), \tag{1}$$

where $\boldsymbol{Z}(k,l)$ denotes the M-channel recordings in the frequency domain, stacked to a $M \times 1$ vector at frequency bin $k = 1, \ldots, K$ and time frame l. $\boldsymbol{S}(k,l)$ and $\boldsymbol{N}(k,l)$ denote the separated multi-channel speech and noise components.[1] For uncorrelated speech and noise signals, the PSD matrix of the input is given as

$$\boldsymbol{\Phi}_{ZZ} = \boldsymbol{\Phi}_{SS} + \boldsymbol{\Phi}_{NN}. \tag{2}$$

[1] For enhanced readability, the frequency and time frame indices will be omitted except where necessary.

Since $\boldsymbol{\Phi}_{SS}$ contains a single speech source, it can be decomposed into the speech PSD Φ_S and the *acoustic transfer functions* (ATFs) $\boldsymbol{A}$ from the speaker to the microphones [12], i.e.

$$\boldsymbol{\Phi}_{SS} = \boldsymbol{A}\boldsymbol{A}^H \Phi_S. \tag{3}$$

3. MULTI-CHANNEL SPEECH ENHANCEMENT

The MSE-optimal multi-channel Wiener filter for estimating the single speaker from the inputs $\boldsymbol{Z}(k,l)$ is given as [13, 14]

$$\begin{aligned}\boldsymbol{W}_{OPT} &= \boldsymbol{\Phi}_{ZZ}^{-1}\boldsymbol{\Phi}_{ZS} \\ &= \left[\boldsymbol{A}\boldsymbol{A}^H\Phi_S + \boldsymbol{\Phi}_{NN}\right]^{-1}\Phi_S\boldsymbol{A} \\ &= \underbrace{\frac{\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}}{\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}}}_{\boldsymbol{W}_{MVDR}} \cdot \underbrace{\frac{\Phi_S}{\Phi_S + \left[\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}\right]^{-1}}}_{G=\frac{\xi}{1+\xi}},\end{aligned} \tag{4}$$

where $\boldsymbol{\Phi}_{ZS}$ is the cross-PSD of $\boldsymbol{Z}(k,l)$ and $\boldsymbol{S}(k,l)$, and the vector $\boldsymbol{W}_{MVDR}$ can be recognized as the MVDR beamformer. G depicts a real-valued, single-channel gain mask. From (4), ξ is given as

$$\xi = \Phi_S\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A} \tag{5}$$

which can be recognized as the SNR at the beamformer output, i.e.

$$\xi = \frac{\boldsymbol{W}_{MVDR}^H\boldsymbol{\Phi}_{SS}\boldsymbol{W}_{MVDR}}{\boldsymbol{W}_{MVDR}^H\boldsymbol{\Phi}_{NN}\boldsymbol{W}_{MVDR}}. \tag{6}$$

3.1. Eigenvector Beamforming

In real-world applications, both the ATFs $\boldsymbol{A}$ and the noise PSD matrix $\boldsymbol{\Phi}_{NN}$ are hard to estimate. The latter might even be ill-conditioned and therefore not invertible. As a consequence, the MVDR beamformer in (4) is difficult to be implemented. Instead, the GSC is widely used [6, 15–18]. The GSC consists of a *steering vector* $\boldsymbol{F}$, a *blocking matrix* $\boldsymbol{B}$, and an *adaptive interference canceller* $\boldsymbol{H}_{AIC}$, i.e.

$$\boldsymbol{W}_{MVDR} \approx \boldsymbol{W}_{GSC} = \boldsymbol{F} - \boldsymbol{B}\boldsymbol{H}_{AIC}. \tag{7}$$

While the GSC avoids the inversion of $\boldsymbol{\Phi}_{NN}$, the steering vector $\boldsymbol{F}$ is still a crucial component, as it directs the beamformer into the direction of the desired speech signal. Obviously, the optimal steering vector would be the ATFs $\boldsymbol{A}$, but they are unknown and hard to estimate in reverberant environments [6]. Eigenvalue decomposition of (3) yields

$$\boldsymbol{\Phi}_{SS} = \boldsymbol{v}_S\boldsymbol{v}_S^H\lambda_S = \boldsymbol{A}\boldsymbol{A}^H\Phi_S, \tag{8}$$

where λ_S and $\boldsymbol{v}_S$ are the principal eigenvalue[2] and eigenvector of $\boldsymbol{\Phi}_{SS}$, respectively. It can be seen that $\boldsymbol{v}_S$ points towards the speech source. The eigenvector includes reverberations and early echoes of the target signal, hence it qualifies as a substitute for the unknown ATFs $\boldsymbol{A}$, and can be used as steering vector $\boldsymbol{F}$. This concept is known as *eigenvector* or *subspace* beamforming [12, 19] where

$$\boldsymbol{F} := \boldsymbol{v}_S. \tag{9}$$

However, $\boldsymbol{\Phi}_{SS}$ cannot be directly observed, but for the purpose of eigenvector decomposition it can be approximated using

$$\hat{\boldsymbol{\Phi}}_{SS}(k,l) = \frac{\sum_{t=1}^{T}\boldsymbol{Z}(k,t)\boldsymbol{Z}^H(k,t)p_{SPP}(k,t)}{\sum_{t=1}^{T}p_{SPP}(k,t)}, \tag{10}$$

[2]Note that $\boldsymbol{\Phi}_{SS}$ is of rank 1 for a single speaker, see (3).

where p_{SPP} is the *speech presence probability* ($0 \leq p_{SPP} \leq 1$), and T is a number of frames during which the dominant eigenvector $\boldsymbol{v}_S$ is assumed to be constant, i.e. the speaker is not moving. Intuitively, using (9), a blocking matrix which satisfies $\boldsymbol{B}^H\boldsymbol{A} \stackrel{!}{=} \boldsymbol{0}_{1\times M}$ is then given by

$$\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{F}\boldsymbol{F}^H = \boldsymbol{I} - \boldsymbol{v}_S\boldsymbol{v}_S^H, \tag{11}$$

where $\boldsymbol{I}$ is the $M \times M$ identity matrix. A similar concept is also used in [20]. The adaptive interference canceller $\boldsymbol{H}_{AIC}$ is usually implemented using an adaptive normalized least mean squares (NLMS) filter [21]. Adaption of this filter has to be stopped while the speaker is active, otherwise target cancellation occurs. Usually this is done using voice activity detection (VAD). However, we used a state-space model [22] to adapt $\boldsymbol{H}_{AIC}$, which does not require a VAD.

Note that the steering vector $\boldsymbol{F}$ and the blocking matrix $\boldsymbol{B}$ depend on the dominant eigenvector $\boldsymbol{v}_S$, hence we refer to this beamformer as *eigenvector GSC* (EV-GSC). Furthermore, $\boldsymbol{v}_S$ depends on the speech presence probability p_{SPP}, see (10). Therefore, the performance of the beamformer depends on an accurate estimate of p_{SPP}.

3.2. Optimal Postfilter

Analogously to (10), the noise PSD matrix $\boldsymbol{\Phi}_{NN}(k,l)$ can be approximated as

$$\hat{\boldsymbol{\Phi}}_{NN}(k,l) = \frac{\sum_{t=1}^{T}\boldsymbol{Z}(k,t)\boldsymbol{Z}^H(k,t)\big(1 - p_{SPP}(k,t)\big)}{\sum_{t=1}^{T}\big(1 - p_{SPP}(k,t)\big)}. \tag{12}$$

Using (6), the SNR at the beamformer output is

$$\xi = \frac{\boldsymbol{W}_{GSC}^H\hat{\boldsymbol{\Phi}}_{SS}\boldsymbol{W}_{GSC}}{\boldsymbol{W}_{GSC}^H\hat{\boldsymbol{\Phi}}_{NN}\boldsymbol{W}_{GSC}} \tag{13}$$

and the postfilter from (4) is given as $G = \frac{\xi}{1+\xi}$. Similar as for the beamformer, the postfilter solely depends on the speech presence probability p_{SPP}.

4. SPEECH MASK ESTIMATION

As demonstrated above, the speech presence probability p_{SPP} is sufficient to construct an optimal multi-channel Wiener filter consisting of our EV-GSC and a postfilter. Therefore, the estimation of p_{SPP} is the key component of our multi-channel speech enhancement system. There are a number of concepts for estimating a speech mask from noisy data, like parameter estimation using a CGMM [11], or neural networks operating on spectrogram data [1,3,9]. Usually, these methods use the signal energy or PSDs as feature vectors, and are therefore highly dependent on the array geometry and statistics of the speech and noise presented in the training data.

However, in a scenario like CHiME4, no reliable assumptions can be made about the signal statistics. The speaker position is unknown, and the background noise is non-stationary and can contain all sorts of sounds from passing-by cars, transient bursts from pneumatic bus doors to human speech. The number of usable microphones can also change, due to microphone failures. Further the array geometry might be unknown, like for the 2 channel track in CHiME4 [10]. Also, the microphones may not be matched. In such situations, the signal power alone is no reliable indicator for speech presence. We observed that the eigenvectors of the PSD matrix $\boldsymbol{\Phi}_{ZZ}$

of the noisy inputs provide a good measure for speaker activity, independent of signal energy and array geometry. We only assume that the speaker is *slowly moving*, and that the noise is *non-stationary*. Eigenvalue decomposition of $\mathbf{\Phi}_{ZZ}$ gives

$$\mathbf{\Phi}_{ZZ} = \sum_{m=1}^{M} \lambda_{Z,m} \boldsymbol{v}_{Z,m} \boldsymbol{v}_{Z,m}^{H}, \tag{14}$$

where $\lambda_{Z,m}$ and $\boldsymbol{v}_{Z,m}$ are the eigenvalues and eigenvectors of $\mathbf{\Phi}_{ZZ}$. We denote $m = 1$ as the dominant eigenvector $\boldsymbol{v}_{Z,1}$. Note that $\lambda_{Z,m}$ corresponds to the signal power, and $\boldsymbol{v}_{Z,m}$ corresponds to the spatial information embedded in the signal.

4.1. Visualization of Eigenvectors

For $M = 6$ channels, the complex-valued eigenvectors $\boldsymbol{v}_{Z,1}(k,l)$ lie on the surface of a 11-dimensional unit sphere[3]. In Figure 1 we show $\boldsymbol{v}_{Z,1}(k,l)$ for 10,000 consecutive frames l from the 'embedded' street recordings. The selected frequency bin k corresponds to $\approx$ 2650Hz. The dots are colored according to $p_{SPP}(k,l)$, which has been calculated from the PSDs of the speech and noise ground truth available for the simulated data of CHiME4, i.e.

$$p_{SPP,\text{true}} = \frac{\text{Tr}\{\mathbf{\Phi}_{SS}\}}{\text{Tr}\{\mathbf{\Phi}_{SS} + \mathbf{\Phi}_{NN}\}}. \tag{15}$$

Using PCA to visualize the first three principal components of $\boldsymbol{v}_{Z,1}$ reveals an interesting structure. It can be seen that the dominant eigenvectors form local clusters if speech is present (red dots). During speech absence they are uniformly distributed over the sphere (blue dots). This clustering indicates that the speaker is indeed *slowly* moving, which will be exploited to estimate p_{SPP}.

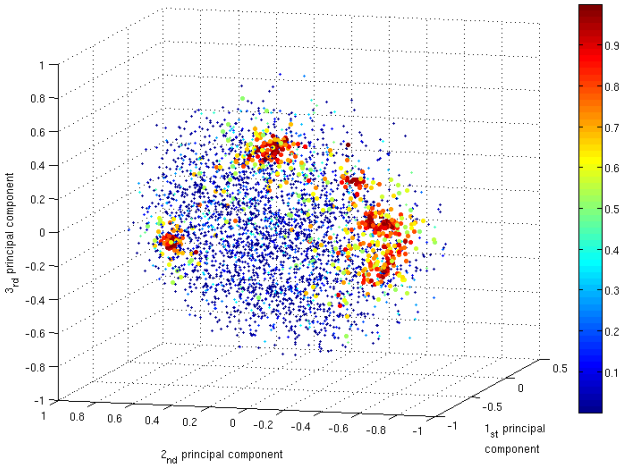

Fig. 1. 3D projection of $\boldsymbol{v}_{Z,1}$ for a single frequency bin over time. The dots are colored according to $p_{SPP,\text{true}}$.

4.2. Kernelized DNN

We use a DNN to learn p_{SPP} from the dominant eigenvector $\boldsymbol{v}_{Z,1}$ of the PSD matrix $\mathbf{\Phi}_{ZZ}$ of the noisy inputs. As we are operating in the frequency domain, a separate kernel for each frequency bin k is required. To introduce some context-sensitivity into our model, we do not use $\boldsymbol{v}_{Z,1}(k,l)$ directly as feature vector, but calculate the cosine distance[4] $x_{i,k}$ between the current eigenvector $\boldsymbol{v}_{Z,1}(k,l)$ at time frame l and the i^{th} most recent frame, i.e.

$$x_{i,k} = \text{Re}\Big\{\boldsymbol{v}_{Z,1}(k,l)^{H} \boldsymbol{v}_{Z,1}(k,l-i)\Big\}. \tag{16}$$

This enables the DNN to exploit the temporal information embedded in the signal. $x_{i,k}$ is stacked to produce a feature vector $\boldsymbol{x}_k$ per kernel k, so that a feature vector covering Δ consecutive frames consists of $\boldsymbol{x}_k = \big[x_{1,k}, x_{2,k}, \cdots, x_{\Delta,k}\big]$. Note that (16) effectively eliminates the number of microphones from the feature vector. Hence, we can apply the same DNN structure to a wide range of multi-channel speech enhancement tasks.

The DNN of each kernel uses a hybrid model with a generative and a discriminative component [2]. The generative component consists of two autoencoder layers, which perform unsupervised clustering of the input data $\boldsymbol{x}_k$. The autoencoder kernels operate independently for each frequency bin. We varied the number of hidden layers and the number of neurons per layer in our experiments, and heuristically determined that 2 hidden layers comprising 20 and 10 neurons are a good compromise between clustering performance and computational complexity. The discriminative component consists of a regression layer which fuses the activations of all autoencoder kernels, in order to exploit information which is distributed across the frequency. The regression layer predicts the K output labels $p_{SPP}(\boldsymbol{x}_k)$). Figure 2 illustrates the kernelized DNN used in our system. Note that we could also use a (bidirectional) long short term memory (B-LSTM), but our kernelized DNN has the advantage of an efficient implementation, and it is easy to train.

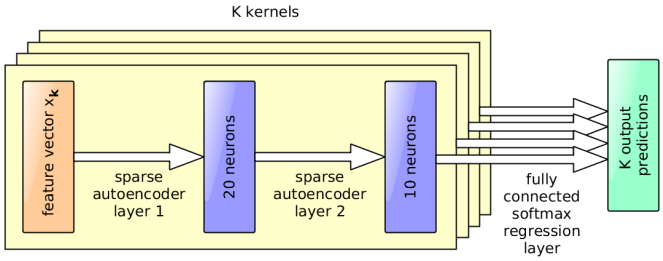

Fig. 2. Kernelized DNN with feature vector $\boldsymbol{x}_k$ and output predictions $p_{SPP}(\boldsymbol{x}_k)$.

4.3. DNN training

We use greedy layer-wise pretraining for the autoencoder kernels [23], and discriminative fine-tuning for the softmax-layer using the true label $p_{SPP,\text{true}}$ from (15). Optimization is done using stochastic gradient descent with ADAM [24]. The autoencoder uses the KL-divergence and weight decay to enforce a sparse representation of the inputs $\boldsymbol{x}_k$. The softmax layer uses the cross entropy between the true and predicted speech presence probability as cost function.

5. RESULTS

We trained our kernelized DNN using the 6-channel training data of the CHiME4 corpus [10], for which the ground truth $p_{SPP,\text{true}}$ is

[3] An m-dimensional complex eigenvector has $2m-1$ non-redundant real-valued dimensons, as the eigenvector can be scaled by an arbitrary complex constant so that one dimension collapses to zero.

[4] Note that the eigenvector is already normalized to 1, i.e. $||\boldsymbol{v}_{Z,1}(k,l)||_2^2 \stackrel{!}{=} 1$

available. The training set comprises 1600 real and 7138 simulated utterances. Then we applied the DNN to the entire 2 and 6-channel corpus consisting of 14658 utterances, which translates roughly into 28 hours of audio data. The DNN outputs the speech presence probability p_{SPP}, which we use to construct the EV-GSC beamformer and postfilter as described in Section 3. For more details on the CHiME4 data the interested reader is refered to [10].

5.1. Speech Mask Accuracy

Figure 3 shows the performance of the DNN for a single utterance from the evaluation set (M04_420C020M_CAF). Panel (a) shows $x_{i=1,k}$ from the feature vector for the DNN, (b) shows the true label calculated with (15), and (c) shows the prediction for p_{SPP} obtained from the softmax regression layer.

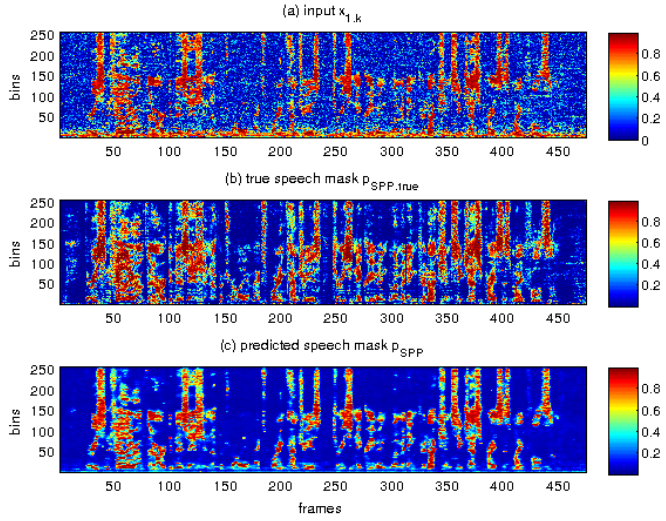

Fig. 3. Speech presence probability mask prediction.

We observe that $x_{i=1,k}$ already shows a high similarity to the true speech presence probability, except for low frequencies and some noise. Due to the fully connected softmax layer, the noise can be almost completely removed, and the prediction accuracy is also good for low frequencies. Table 1 reports the prediction error for p_{SPP}[5] for the 2 and 6 channel data of CHiME4 [10], and various feature vector lengths Δ. Using a feature vector with more than 5 consecutive frames gives no significant performance improvement, hence we select $\Delta = 5$ to be a reasonable trade-off between accuracy and computational complexity.

Scenario	Train	Dev	Eval
2ch, $\Delta = 3$	15.46	15.85	16.58
2ch, $\Delta = 5$	15.08	15.61	16.17
2ch, $\Delta = 7$	14.89	15.32	16.02
6ch, $\Delta = 3$	11.16	11.69	12.24
6ch, $\Delta = 5$	10.74	11.41	11.85
6ch, $\Delta = 7$	10.55	11.28	11.74

Table 1. Prediction error for p_{SPP} in %.

[5]The prediction error is the average over all time-frequency bins of $|p_{SPP} - p_{SPP,\text{true}}|$.

5.2. Perceptual Speech Quality

With the predicted speech mask p_{SPP}, we construct the EV-GSC beamformer from Section 3.1. We use the *Perceptual Evaluation methods for Audio Source Separation* (PEASS) Toolkit [25, 26] to evaluate the performance of our multi-channel speech enhancement system, and report the *Overall Perceptual Score* (OPS) and PESQ [27] values. Tables 2 and 3 give a comparison of our system (EV-GSC) against the CHiME4-baseline enhancement system using the BeamformIt-toolkit [10], our GSC with steering vector estimation and an *adaptive blocking matrix* (ABM) [7], and the front-end of the best CHiME3 system [11], which uses a complex gaussian mixture model (CGMM-EM) to estimate the speech and noise PSD matrices. The model parameters are estimated with an EM algorithm, and the posterior probability is used as speech presence probability.

It can be seen that our approach (EV-GSC) outperforms the CHiME4 baseline, the GSC with ABM, and the CGMM-EM systems in terms of PESQ and OPS on the simulated (simu) and real (real) data set for 6-channels (6ch). Even in the 2-channel case (2ch) we obtain competitive results. In this case, the two channels are randomly selected from the 6-channels, i.e. the array geometry changes randomly.

Method	Data set	Train	Dev	Eval
CHiME4 baseline	simu	1.35	1.31	1.26
(BeamformIt), 5ch [10]	real	1.35	1.28	1.37
GSC with ABM and	simu	1.98	1.69	1.63
and postfilter, 6ch [7]	real	1.51	1.39	1.44
CGMM-EM with MVDR	simu	1.79	1.59	1.51
and postfilter, 6ch [11]	real	1.53	1.41	1.44
EV-GSC and postfilter,	simu	**2.04**	**1.89**	**1.86**
6ch, $\Delta = 5$	real	**1.72**	**1.74**	**1.63**
EV-GSC and postfilter,	simu	1.68	1.61	1.58
2ch, $\Delta = 5$	real	1.55	1.43	1.54

Table 2. PESQ scores.

Method	Data set	Train	Dev	Eval
CHiME4 baseline	simu	33.11	34.73	31.46
(BeamformIt), 5ch [10]	real	29.97	36.45	36.74
GSC with ABM and	simu	56.08	44.82	44.48
and postfilter, 6ch [7]	real	47.18	44.90	36.96
CGMM-EM with MVDR	simu	52.15	43.02	40.59
and postfilter, 6ch [11]	real	44.95	41.89	36.87
EV-GSC and postfilter,	simu	**59.09**	**48.32**	**48.64**
6ch, $\Delta = 5$	real	**52.34**	**46.09**	**44.16**
EV-GSC and postfilter,	simu	47.32	40.97	40.75
2ch, $\Delta = 5$	real	43.43	42.83	39.82

Table 3. OPS scores.

6. CONCLUSION AND FUTURE WORK

In this paper, we have shown the importance of the speech presence probability mask, which is used to construct an optimal multi-channel Wiener filter followed by a single-channel postfilter. Further, we presented a kernelized DNN to estimate this speech presence probability mask. To prevent the DNN from learning specific speaker and noise types, we used the cosine distance between the dominant eigenvectors of consecutive frames of the PSD of the noisy speech as input feature. Finally, we compared our system against three state-of-the-art approaches, and evaluate the perceptual speech quality and classification error. Future work includes a in-depth evaluation of the DNN being used and performance comparison against B-LSTMs. Furthermore, the relationship between the eigenvectors and the speech presence probability mask is investigated.

7. REFERENCES

[1] Y. Wang and D. Wang, "Cocktail party processing via structured prediction," in *Neural Information Processing Systems (NIPS)*, 2012, pp. 224–232.

[2] M. Zöhrer, R. Peharz, and F. Pernkopf, "Representation learning for single-channel source separation and bandwidth extension," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2398–2409, 2015.

[3] F. Weninger, J. Le Roux, J. R. Hershey, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. GlobalSIP Symposium on Machine Learning Applications in Speech Processing*, Dec. 2014.

[4] L. Deng, M.L. Seltzer, D. Yu, A. Acero, A. Mohamed, and G.E. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder.," in *Interspeech*, 2010, pp. 1692–1695.

[5] M. Zöhrer and F. Pernkopf, "Single channel source separation with general stochastic networks," in *Interspeech*, 2014.

[6] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, Springer, Berlin–Heidelberg–New York, 2008.

[7] L. Pfeifenberger, T. Schrank, M. Zöhrer, M. Hagmüller, and F. Pernkopf, "Multi-channel speech processing architectures for noise robust speech recognition: 3rd chime challenge results," in *Proc. IEEE ASRU*, 2015.

[8] L. Pfeifenberger and F. Pernkopf, "Blind source extraction based on a direction-dependent a-priori SNR," in *Interspeech*, May 2014.

[9] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE ICASSP*, 2016.

[10] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE 2015 Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015.

[11] Higuchi T., Ito N., Yoshioka T., and Nakatani T., "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 5210–5214, Mar. 2016.

[12] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer, Berlin–Heidelberg–New York, 2008.

[13] P. Vary and R. Martin, *Digital Speech Transmission*, Wiley, West Sussex, 2006.

[14] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing*, Springer, Berlin–Heidelberg–New York, 2006.

[15] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, Oct. 1999.

[16] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, Aug. 2001.

[17] W. Herbordt and W. Kellermann, "Analysis of blocking matrices for generalized sidelobe cancellers for non-stationary broadband signals," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, May 2002.

[18] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 6, Nov. 2004.

[19] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, July 2007.

[20] E. Warsitz, A. Krueger, and R. Haeb-Umbach, "Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 73–76, May 2008.

[21] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, New Jersey, 4th edition, 2002.

[22] F. Kuech, E. Mabande, and G. Enzner, "State-space architecture of the partitioned-block-based acoustic echo controller," *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2014.

[23] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Neural Information Processing Systems (NIPS)*, 2007.

[24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference for Learning Representations, San Diego, 2015*, July 2015.

[25] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," *10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 430–437, 2012.

[26] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, Sept. 2011.

[27] "ITU-T recommendation P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2000.